CHAPTER 3

STATISTICAL CLASSIFICATION METHODS

The employees of the ISB use the WTV statements of an insurance company to assess its risk exposure (see chapter 1, section 5). Analyzing statements is an important tool to assess the risk exposure of all kinds of companies. Such an analysis is usually conducted by calculating a number of financial ratios. An important part of the analysis is to compare the financial ratios to specific norms on a univariate and/or multivariate basis. Thus, a company can be classified into one of several risk categories, depending on its score(s) on the univariate and/or multivariate financial ratio test(s). This chapter will focus on a number of statistical methods which can be used for financial ratio analysis. Given the number of succesfull applications in the past, statistical methods are an obvious candidate for inclusion in N.E.W.S.. Two other types of methods which are considered for inclusion in N.E.W.S., *i.e.* neural networks and expert systems, will be discussed in chapters 4 and 5. Both statistical methods and neural networks are mathematical models, while expert systems are rule-based models.

A particular application of company assessment is the identification of financially distressed companies, that is, the identification of companies that are likely to go bankrupt if no remedial actions are taken. A classification model, based on a univariate and/or multivariate comparison of financial ratios, which can identify financially distressed companies in time, can therefore be used as an early warning system by supervisors and other stakeholders.

In the next section, a number of techniques are presented, which can be used to obtain estimates of the true error rate of a classification model. These error rate estimations can be used to compare the adequacy of different classification models for a certain application, like identifying financial distress. In the remaining sections, a survey will be given of the most common approaches to identify financial distress, *i.e.* univariate ratio analysis (section 3.2), dichotomous multiple discriminant analysis (3.3.1), and dichotomous logit and probit analysis (3.3.2), and the polytomous extensions of multiple discriminant analysis, and logit and probit analysis (3.4). The polytomous extensions, which, to my knowledge, have not been used for identifying financial distress before, are presented because they play an important role in this research project. For each approach a number

of references to earlier applications, mainly aiming at identifying financial distress, will be given.

3.1 ERROR RATE ESTIMATION

The objective of deriving classification rules from sample data is to classify and successfully predict new data. The most commonly used measure of success or failure is a classifier's error rate, which is the ratio of the number of errors to the number of cases (or observations). The *true* error rate is statistically defined as the error rate of the classifier on a large number of new cases, which in the limit converge to the actual population distribution. In other words, the true error rate is defined as the error rate of the classifier if it had been tested on the true distribution of cases in the population, which can be empirically approximated by a very large number of new cases gathered independently from the cases used to design the classifier. In the real world, the number of sample cases available is finite and typically relatively small. The major question is then whether it is possible to extrapolate from empirical error rates calculated from small sample results to the true error rate. There are several techniques, some far better than others, which are used to determine estimates of the true error rate. Some of the most popular techniques will be discussed below. See [Weiss & Kulikowski, 1991, pp. 17-49] for a more extensive treatment.

A "naive" technique is to use the error rate of the classifier for the sample cases that were used to design or build the classifier. This is called the apparent, resubstitution or reclassification error rate. For most types of classifiers, this error rate is a poor estimator of future performance. The true error rate is almost always higher.

Instead of using all available cases to build the classifier and to estimate the true error rate, the cases can be divided into two groups: a training set, which is used to design the classifier, and a testing set (or holdout sample) to test the classifier. With respect to the single train-and-test (or holdout) method, a fixed percentage of cases is used for training, and the remainder for testing. The usual proportions are approximately a 2/3, 1/3 (*i.e.* 2:1) split. The holdout sample error rate estimate is far stronger than the apparent error rate. With a large number of sample cases, it is a very reasonable approach. With moderately sized samples, the holdout method usually leaves one with insufficient training or testing cases. Using a typical 2/3 and 1/3 partition, the holdout estimate is a relatively pessimistic estimate of the true error rate for small or moderately-sized samples. In such

case, resampling methods provide better estimates of the true error rate. These methods are variations of the (single) train-and-test method: instead of a single random partition, which can be misleading for small or moderately-sized samples, multiple train-and-test experiments are performed.

When multiple random train-and-test experiments are performed, a new classifier is generated from each training sample. The estimated error rate for the classifier based on all sample cases is the average of the error rates for classifiers derived from the independently and randomly generated test partitions. A special case of resampling is known as leaving-one-out, introduced by Lachenbruch [1967]. This technique is especially attractive when relatively small sample sizes are available. For a given sample size n, a classifier is generated using $(n-1)$ cases and is tested on the single remaining case. This is repeated n times, each time designing a classifier by leaving-one-out. Thus, each case in the sample is used as a test case, and each time nearly all cases are used to design a classifier. The error rate estimator (*i.e.* the number of errors on the separate test cases divided by n) is an almost unbiased estimator of the true error rate of the classifier based on the full sample of size n.

Although leaving-one-out is a preferred technique, with large sample sizes it may be quite expensive from a computational perspective. However, as the sample size grows, other train-and-test methods improve their accuracy in estimating error rates. The leaving-one-out error estimation technique is a special case of the general class of cross-validation error estimation methods. In k-fold cross-validation, the cases are randomly divided into k mutually exclusive test partitions of approximately equal size. The separate cases excluded from each test partition are used independently for training, and the resulting classifier is tested on the corresponding test partition. The average error rates over all k partitions is the cross-validated error rate. For fivefold cross-validation, for instance, five iterations are needed with a training set containing 80% of the total sample and a testing set with the remaining 20%. The main advantage of cross-validations is that all cases in the sample available are used for testing, and almost all cases are used for training the classifier as well.

Weiss and Kulikowski [1991, p. 38] provide the following guideline concerning the particular resampling techniques to be used: for sample sizes larger than 100, either tenfold cross-validation or leaving-one-out is acceptable. Tenfold cross-validation is far less expensive from a computation point of view than leaving-one-out, and it can be readily used for sample sizes of several hundreds. For sample sizes less than 100 they advocate the use of the leaving-one-out procedure. In this study, a 1992 training set with 195 observations is used (see chapter 6).

To test the logit model, both tenfold cross-validation and a holdout sample with 1993 data are used.

3.2 UNIVARIATE RATIO ANALYSIS

Early studies on financial distress were usually based on analyses of individual financial ratios or groups of financial ratios on a univariate basis.

Beaver [1966] employed dichotomous classification tests and likelihood analyses based on individual financial ratios to classify industrial firms as bankrupt or nonbankrupt one, two, or three years before actual bankruptcy took place. By trial and error, a cutoff point that minimizes the number of misclassifications is determined for each ratio. Thus, for each ratio, a prediction can be made that firms with ratio values higher and lower than the cutoff point can be divided into two groups.

In order to assess likelihood ratios[1], Beaver prepared histograms of the dispersion of the ratios for the two groups. The divergence of the distributions of the ratios for the two groups was visually apparent as the time until bankruptcy shortened. Although the likelihood ratios were not calculated directly, the probabilities of the ratios having certain values if the firm would be in the failed or the nonfailed group could be estimated from the heights of the distributions.

Although Beaver's predictors performed fairly well, the main difficulty with his approach is that classification can take place for only one ratio at a time. It is possible to find conflicting classifications of any given firm according to various ratios [Zavgren, 1983, p. 10]. For a more comprehensive discussion of the work of Beaver, see Zavgren [1983, pp. 3-10].

The IRIS (Insurance Regulatory Information System) ratio system, promulgated by the NAIC (National Association of Insurance Commissioners) and introduced in the early 1970s, uses a number of financial ratios to support the assessment of the financial strength of an insurance company. This system consists of twelve ratio tests for life & health insurers and eleven ratio tests for property & liability insurers. A company will be assigned a high priority for further investigation if four or more ratios fail to meet their respective norms[2].

---

[1] Likelihood ratios indicate the frequency with which the values of the ratios occur at certain intervals (conditional on the firm being classified as failed or nonfailed).

[2] For a Dutch description of the IRIS system, see Sprangers [1981].

The IRIS system has been criticized for its inability to consider interdependence among ratios, the seemingly *ad hoc* ranges of some of the ratios, and failing to actually provide early warnings of financial distress [Harrington & Nelson, 1986, p. 585]. Furthermore, Eck [1982, p. 448] criticized the heavy reliance on loss reserves of the property & liability system. The loss reserve or the surplus is included in nine of the eleven tests. Management would be able to alter the loss reserve figures relatively easily, and the test results would be satisfactory and could "pass" nine of the eleven tests as a consequence.

3.3 DICHOTOMOUS MULTIVARIATE RATIO ANALYSIS

The financial status of a firm is actually multidimensional, and no single ratio is able to capture those dimensions [Zavgren, 1983, p. 10]. In this field, therefore, the emphasis gradually shifted to multivariate techniques. With multivariate analysis, the predictive ability is jointly analyzed for several financial ratios. Within the whole set of multivariate techniques, this section will focus on the case of a binary dependent variable. Polytomous models, *i.e.* models where the dependent variable can take on more than two states, will be the subject of section 3.4.

3.3.1 Multiple Discriminant Analysis

The pioneering study in the application of multivariate statistical analyses to the prediction of financial distress has been conducted by Altman [1968]. Altman, and after him many more researchers, applied Multiple Discriminant Analysis (MDA) in order to predict corporate bankruptcy. The description of MDA given below is taken mainly from Altman *et al.* [1981, pp. 33-52].

MDA was originally developed by Fisher [1936]. It evolved as a variant and multivariate extension of the univariate analysis of variance techniques. MDA assumes known, identifiable, mutually exclusive groups. A sample of observations is drawn from each group, where each observation is described by measurements of a set of variables. Under certain assumptions, a linear function is estimated, which optimally discriminates between the two groups by maximizing the ratio of the between-groups and the within-groups variance. By making assumptions concerning the distribution of the variables within each group, one can address such issues as (1) testing for differences in variable mean vectors and/or covari-

ance structures across groups, and (2) constructing schemes to estimate ex-ante probabilities of group membership for future or uncertain observations given information on the independent variables only.

We will assume two identifiable populations, representing for instance bankrupt and nonbankrupt firms. Population (group) membership will be presented by variable y where $y_n = 2$ if the nth firm is bankrupt (group 2), and by $y_n = 1$ in any other case (group 1). Furthermore, it will be assumed that each firm is also characterized by a vector X of m explanatory variables $x^{(i)}$. Thus, for the nth observation, X is represented as an m length column vector

$$\mathrm{X_n} = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(m)})'.$$

These may be financial variables (*s.a.* the net income/total assets ratio or firm size) whose distribution is assumed to be related to bankruptcy. Within each group, X is assumed to be distributed with a multivariate normal distribution. Thus,

$$\mathrm{X_n} \sim N(\mu_1, \Sigma_1) \qquad \text{given } y_n = 1,$$

and

$$\mathrm{X_n} \sim N(\mu_2, \Sigma_2) \qquad \text{given } y_n = 2,$$

μ_1 and μ_2 being m length mean vectors and Σ_1 and Σ_2 ($m \times m$) covariance matrices. Thus, the groups are assumed to be given. Consequently, given group membership, the distribution of the X variables is determined.

Given the observed characteristics of the nth firm $\mathrm{X_n}$, the likelihood that a firm with these characteristics would be generated if the prevailing group were nonbankrupt firms is the multivariate normal density function with parameters μ_1 and Σ_1, $f_1(\mathrm{X_n})$. Similarly, the likelihood of observing $\mathrm{X_n}$, given the fact that the generating group refers to bankrupt firms, is the multivariate normal density function with parameters μ_2 and Σ_2, $f_2(\mathrm{X_n})$. Classification probabilities and rules are constructed by comparing the two group likelihood functions.

Suppose, for example, the researcher perceives two types of costs of misclassification, C(1|2) and C(2|1), C(i|j) being the cost of classifying an observation as group i when it actually belongs to group j. Furthermore, let us assume we want to construct a classification scheme that will minimize expected costs from misclassification for a random draw of the population. In such case, the following classification rule should be applied: an observation with characteristics $\mathrm{X_n}$ should be assigned to group 1 if

$$\frac{f_1(X_n)}{f_2(X_n)} \geq \frac{C(1|2)\pi_2}{C(2|1)\pi_1} ,$$

and in any other case to group 2. π_1 is the *a priori* probability of an observation taken from group 1, and π_2 is the *a priori* probability of an observation taken from group 2 ($\pi_1 + \pi_2 = 1$). They can be interpreted as the relative sizes of the two populations. If the costs of misclassification are equal, *i.e.* $C(1|2) = C(2|1)$, the classification rule is identical to assigning observations to the group with the highest probability given its vector of characteristics, X_n.

If the group covariance matrices are unequal, *i.e.* $\Sigma_1 \neq \Sigma_2$, the classification rule can be expressed as a quadratic function in X_n (see Altman *et al.* [1981, p. 40]). In most applications, however, the additional assumption is made of equal group covariance matrices, *i.e.* $\Sigma_1 = \Sigma_2$. In such case, the classification rule is reduced to a linear function: an observation with characteristics X_n should be assigned to group 1 if

$$X_n'\gamma - \alpha \geq \ln[C(1|2)\pi_2/C(2|1)\pi_1] ,$$

with

$$\gamma = \Sigma^{-1} (\mu_1 - \mu_2) ,$$

$$\alpha = (\mu_1 + \mu_2)'\gamma/2 ,$$

and in any other case to group 2.

Vector γ can be considered as weights of the population *discriminant function*, and the inner product $X_n'\gamma$ as the *discriminant score* for a particular observation. The linear classification rule implies a simple classification heuristic. The discriminant score, computed as a linear weighting of X_n, is compared to a preset cutoff, $\alpha + \ln[C(1|2)\pi_2/C(2|1)\pi_1]$. If the score is higher than the cutoff, the observation is assigned to group 1, otherwise it is assigned to group 2.

If the population parameters are unknown, which is usually the case for practical applications, sample estimates $\{\bar{X}_1, \bar{X}_2, S\}$ can be substituted for $\{\mu_1, \mu_2, \Sigma\}$ in the classification rule. Assume a random sample of size $N_1 + N_2 = N$, N_1 being known as belonging to group 1 and N_2 being known as belonging to group 2. Defining X_{1n} as observations from group 1 and X_{2n} as observations from

group 2, the sample estimates can be calculated as follows:

$$\bar{X}_1 = N_1^{-1} \sum_{n=1}^{N_1} X_{1n},$$

$$\bar{X}_2 = N_2^{-1} \sum_{n=1}^{N_2} X_{2n},$$

$$S = \frac{\sum_{n=1}^{N_1} (X_{1n} - \bar{X}_1)(X_{1n} - \bar{X}_1)' + \sum_{n=1}^{N_2} (X_{2n} - \bar{X}_2)(X_{2n} - \bar{X}_2)'}{N - 2}.$$

Assuming random sampling with known groups, rules constructed from sample parameters will be consistent estimators of the true population rules. The sample estimate of the discriminant function vector γ equals:

$$b = S^{-1}(\bar{X}_1 - \bar{X}_2).$$

The sample discriminant vector b maximizes the ratio of the between-groups sample variance of X_n to the within-groups variance holding the sample deviation sums of squares of X_n constant [Altman *et al.*, 1981, p. 52].

Most applications of MDA suffer from statistical or methodological problems. With respect to many applications, MDA is an inappropriate statistical technique. Usually, not enough thought is given to matching the statistics to the underlying theoretical problem. Some of the statistical problems that are most common will be treated in the following paragraphs. For a more extensive treatment, the reader is referred to Altman *et al.* [1981, pp. 119-165].

First, the key assumptions of the classical linear discriminant analysis model are that (1) the variables describing the members of the group are multivariate normally distributed within each group, (2) the group covariances are equal across all groups, and (3) the groups are discrete, mutually exclusive, and identifiable. In practice, deviations from the normality assumption in economics, finance, and insurance tend to be the rule rather than the exception. The tests of significance

and estimated classification error rates may be biased when the normality assumption is violated. Hence, it is of interest to determine whether the assumption holds and what effects its relaxation may have on the statistical test and the ensuing classification. In the applied literature, the problem of testing for the appropriateness of the distributional assumption is largely ignored.

The equality of group covariances is usually not tested. Depending on the sample sizes, number of variables, and differences in the covariances, use of linear classification rules when quadratic rules are indicated may have dramatic effects on the classification results, and in particular on individual assignments and respective group error rates. The test for the equality of the covariances matrices should precede both the test for the equality of group means and the estimation of classification errors.

A third and implicit assumption of MDA is the fact that the groups investigated are distinct and discrete. It is assumed that an observation is a member of only one group. Furthermore, in the parameter estimation process, it is assumed that it is known to which group each sample observation belongs. However, in the applied literature, numerous examples occur that violate these assumptions or that exhibit related problems with the groups that limit the practical usefulness of the classification results. Perhaps the most extreme problem concerning the definition of the groups occurs when an inherently continuous variable is arbitrarily segmented and used as a basis to form groups.

One of the most widely misunderstood aspects of MDA relates to the problem of determining the relative importance or contribution of individual variables to the analysis. Unlike the case with the classical linear regression model, where the normal equations define a unique set of coefficients as a solution, discriminant function coefficients are not uniquely defined; only the ratios between the coefficients are. Therefore, it is difficult to isolate and test behavioral hypotheses concerning the role of individual variables in a model.

The standard MDA classification rules incorporate *a priori* probabilities to account for the relative occurrence of observations in different populations and misclassification costs to adjust for the fact that some classification errors may be more serious than others. The importance of the *a priori* probabilities and/or costs of misclassification have been grossly overlooked. It is not uncommon for many applied studies to employ the techniques without using random samples from the populations or without mentioning any of the *a priori* probabilities used. Many others simply assert that equal *a priori* probability and costs have been assumed, seemingly ignoring the effect this may have on classification performance.

Applications

MDA has been widely used to classify bankrupt and nonbankrupt firms. Trieschmann and Pinches [1973] were among the first to use MDA to predict financial distress for property & liability (p-l) insurers. Their sample consisted of 52 p-l insurers: 26 insurers who entered into liquidation, receivership, conservatorship, or rehabilitation during the period 1966-1971, and 26 randomly selected solvent insurers. For the development of the model the authors used data referring to two years before actual financial distress. Their model, which included six independent variables, could correctly classify 49 firms (94.2%). One solvent firm was classified as being distressed while two of the distressed firms were classified as belonging to the solvent group. Prior probabilities and costs of misclassification were not taken into account.

In Pinches & Trieschmann [1974], the same authors compared the performance of their MDA model with a univariate model for the same sample as in Trieschmann & Pinches [1973]. Three univariate models were tested: a single variable, a three variable, and a five variable model. The five variable model performed best among the univariate models. Firms would be classified as distressed when at least three of the ratios indicated financial distress (*i.e.* were below or over a certain threshold). This model correctly classified 45 firms (86.5%). All solvent firms and 19 of the 26 distressed firms were classified correctly. The MDA model performed better than the single and the three variable univariate models. The differences between the MDA model and the five variable univariate model were not very significant but they still indicated the relative superiority of the multivariate approach for insurance company surveillance [Pinches & Trieschmann, 1974, p. 574].

In Pinches & Trieschmann [1977], the authors tested the independent variables of their previous study for multivariate normality and equal covariance matrices. They concluded that the variables were not multivariate normally distributed, and that the two covariance matrices were not equal. Therefore, since the assumptions of MDA were not met, the classification results of the model might have been biased.

The authors also tested the classificatory power of their MDA model by means of the leaving-one-out technique (see section 3.1). This technique gives an almost unbiased estimate of the probability of misclassification of a model. The technique of using the original sample, used to estimate the classification rule, which was followed by the authors in their previous studies, gives an overall optimistic, biased estimate of how well the rule would perform in the population (see section 3.1). Applying the leaving-one-out technique, the percentage of correct predictions fell to 86.5% (was 94.2%), because 5 out of 26 distressed firms

and 2 out of 26 solvent firms were classified incorrectly.

Harmelink [1974] used MDA to predict a decline or maintenance in Best's general policyholders' ratings for property-liability companies. A company is labelled as unsuccessful if its rating declined from A+ or A to B+ or lower. The sample consisted of 55 unsuccessful companies from the 1960-1970 period. For each unsuccessful company in the sample, a successful company (one that maintained the A+ or A rating) of approximately the same asset size was selected. Thus, the data consisted of a matched sample of 55 successful and 55 unsuccessful companies. Harmelink derived several hundreds of discriminant functions from a set of seven variables for different years and different groups of sample data. On a holdout sample, the best function had an overall predictive ability of 78% one year before a decline set in. The percentage of unsuccessful companies that were classified correctly thus equaled 67%, while 93% of the successful companies were classified correctly.

Sinkey [1975] employs quadratic MDA to identify financial characteristics that distinguish between problem and nonproblem banks. Problem banks are identified by the FDIC (Federal Deposit Insurance Corporation) during bank examinations. The sample consists of 110 problem banks from the period 1972-1973. Each problem bank was matched with a nonproblem bank. By means of the leaving-one-out technique, the quadratic classification rule with ten independent variables correctly classified 72% of the problem banks and 79% of the nonproblem banks, using 1972 data. Given the high degree of group overlap found by the author, the classification results were better than expected.

Pettway and Sinkey [1980] used a dual screening technique to schedule bank examinations, namely screening based on accounting information and screening based on market information. The accounting screen is based on MDA with two independent variables. The sample consisted of 33 banks that failed between 1970 and 1975. Each failed bank was matched with a nonfailed bank. In a holdout sample test, the model correctly identified 15 of the 16 future failures one year prior to the failure and 14 of 16 two years before the failure. The market screen is based on two different tests of the portfolio returns on actively traded securities of the banks. With respect to six major banks that failed and did have sufficient trading of their securities, both tests identified five of the problem banks before the supervisory agency came into action. As for one problem bank, one of the tests did not indicate problems before supervisory action was taken.

Ambrose and Seward [1988] incorporated Best's general policyholder rating

and financial size rating[3] with financial variables of property-liability insurers. The sample consisted of 29 insurers who failed during 1969-1983, and 29 nonbankrupt insurers who were selected to match the bankrupt ones. Three models were tested. With only Best's ratings as independent variables, 79% of the firms was correctly classified (83% for bankrupt and 76% for nonbankrupt firms) by a four-variable model. With only financial ratios, 85% was classified correctly (83% for bankrupt and 86% for nonbankrupt firms). All percentages were calculated with the leaving-one-out technique. The performances of Best's ratings and of financial ratios are statistically equivalent.

Ambrose and Seward finally used a two-stage technique which first separated the sample into a bankrupt and an nonbankrupt group applying MDA based on Best's ratios. Subsequently, the classification table for Best's ratings was used to estimate the probability that a nonbankrupt firm had been classified correctly. This value would be the prior probability of being a nonbankrupt firm. Next, this prior probability was used in the discriminant function as the probability of the insurer being nonbankrupt rather than the default prior of 50%. The same procedure was repeated for the bankrupt group. Prior probabilities for the bankrupt and the nonbankrupt groups following from the first MDA model, *i.e.* based on Best's ratios, are not the same, which implies that using the same set of financial ratios results in different coefficient vectors. This model correctly predicted 88% of the companies (97% for bankrupt and 79% for nonbankrupt companies).

3.3.2 Logit Analysis

By far the most common empirical model form used in economics and finance is regression. Both because of its comprehensibility and ease of use, the regression framework is particularly attractive. It is natural, therefore, to seek ways of casting binary dependent variable models as regressions. However, the regression approach has got some conceptual and computational problems when the dependent variable is binary (see for instance Altman *et al.* [1981]). Of particular concern is the possibility that the regression model can lead to probability predictions below zero or larger than one. A natural response would be to seek out transformations that would restrict probability predictions to the zero-one interval. Particularly desirable transformations would be those where probabilities never actually

[3] Best's general policyholder rating grades the overall health of the insurer. The financial size rating groups firms on the basis of their amount of surplus.

equal zero or one. This is consistent with the view that even for extreme outliers no event can be predicted with the absolute certainty implied by a probability of one or zero.

In the context of a bankruptcy model, y_n is defined as a dependent variable that takes on only two values: 1 if the nth firm goes bankrupt, and 0 in any other case. We want to model the conditional probability of bankruptcy (or conditional expectation of y_n) as a function of a vector of m explanatory variables $X_n = (x_n^{(1)}, x_n^{(2)},..., x_n^{(m)})'$:

$$P_n = E(y_n|X_n),$$

with P_n restricted to the zero-one interval.

A class of transformations that restrict P_n to the zero-one interval and that have the additional attraction of being monotonically increasing (or decreasing) in X_n are cumulative distribution functions, or

$$F(X_n'\beta) = \int_{-\infty}^{X_n'\beta} f(z)\,dz,$$

where F is a cumulative distribution function and f the density function for a given probability function. Note that if one of the variables always equals 1, the linear weighting $X_n'\beta$ can include a constant term. Two particularly attractive cumulative distribution functions are the standard normal and the logistic functions. If the cumulative standard normal distribution function is selected, this is referred to as a **probit model**. In the following, our attention will, however, mainly be focused on the cumulative logistic function.

Suppose that the probability of bankruptcy, given a firms vector of variables X_n, is given by $P_n = F(X_n'\beta)$ where F is the cumulative standard logistic distribution function

$$F(X_n'\beta) = \int_{-\infty}^{X_n'\beta} f(z)\,dz = \frac{1}{[1 + \exp(-X_n'\beta)]},$$

with $f(z)$, the standard logistic density function, equaling

$$f(z) = \frac{\exp(z)}{[1 + \exp(z)]^2}.$$

This model is termed the **logit model**. The logit probability P_n is constrained to the zero-one interval and becomes asymptotic to these values as X_n shifts to $-\infty$ and ∞. The partial derivatives of the conditional probability with respect to X_n (or the change in probability for a given change in X_n) depend on the probability and are smallest in the tails and largest at $P_n = 0.5$. This is a desirable feature, because it is reasonable to think that a given change in an independent variable will have a smaller effect on firms with either very high or very low probabilities than on firms with a probability close to 0.5.

The shape of the standard logistic and the standard normal distribution and density functions do not differ very much: like the normal distribution, the standard logistic distribution is symmetric with mean, mode, and median at zero. However, it has a variance of 3.29 and a standard deviation of 1.81, compared to a variance and a standard deviation of one for the standard normal distribution. Thus, the logistic distribution has more mass shifted to the tails. Another difference between the normal and logistic distributions is that, unlike the normal distribution, the logistic distribution has a "closed form" expression for its cumulative distribution. This makes the logistic distribution considerably more computationally tractable. In other respects, the logit model is quite similar to the probit model. Both imply probabilities of 0.5 when $X_n'\beta = 0$.

An important transformation of P_n is the **logit transformation**. This transformation is defined as follows:

$$g(X_n) = \ln[\frac{P_n}{1 - P_n}] = X_n' \beta \ .$$

The ratio $P_n/(1-P_n)$ is called the odds of the presented outcome (*i.e.* the odds of bankruptcy for the nth firm). The log of the odds, as defined above, is called the logit.

The importance of this transformation is that $g(X_n)$ has many of the desirable properties of a linear regression model. The logit, $g(X_n)$, is linear in its parameters, may be continuous, and may range from $-\infty$ to ∞, depending on the range of the elements of X_n. Furthermore, the slope coefficient $\beta^{(i)}$ represents the change in the logit for a change of one unit in the independent variable $x_n^{(i)}$.

The use of the probit and logit models can be justified for several reasons. Attractive features of the models are that they are monotonic, bounded in the zero-one interval, and asymptotic to zero or one in the tails. In addition to this, a number of underlying structural equations give rise to probit or logit models. A commonly used specification [Altman *et al.*, 1981, pp. 18-19] is to assume that

there is a relationship between the binary dependent variable y_n and an index function of a vector of independent variables $I(X_n)$. The index is assumed to be a linear function of the x's:

$$I(X_n) = X_n' \beta = \sum_{i=1}^{m} x_n^{(i)} \beta^{(i)} .$$

The dependent variable is assumed to be determined by

$$y_n = \begin{cases} 1 & \text{if } I(X_n) \geq \epsilon_n \\ 0 & \text{if } I(X_n) < \epsilon_n \end{cases},$$

where ϵ_n is a stochastic, unobserved error term which is normally distributed $N(0,1)$ in the probit model, or distributed as a standard logistic in the logit model. The index could, for example, be considered a measure of vulnerability to fail [Martin, 1977]. The stochastic term is thought of as a "tolerance of vulnerability." Thus, companies are assumed to fail whenever their vulnerability exceeds their tolerance. Note that the explanatory variables are not bound to any particular distribution, which is in contrast to MDA in which the explanatory variables are assumed to be distributed on the basis of a multivariate normal distribution. For yet another specification that leads to the probit or logit model based on utility maximization, see Altman *et al.* [1981, pp. 19-23]. For this specification as well, no assumptions are made concerning the distribution of the independent variables, but only about the error terms. For a more extensive treatment of logistic regression, the reader is referred to the following texts: Hosmer & Lemeshow [1989], Altman *et al.* [1981], and Maddala [1983].

Although the logit model may lead to similar predictions, the assumptions are markedly different from those related to the MDA model. The logit model evolved from the traditional linear regression model. With respect to the logit model, the line of causality runs from the exogenously determined X variables and stochastic errors ϵ to a random variable Y, which takes on only two discrete values. This relationship is assumed to be deterministic in that if the Xs and the ϵs were known, Y could be solved. Thus, bankruptcy is an endogenous variable determined by an explicit, structural model. MDA evolved as a variant and multivariate extension of univariate analysis of variance techniques. MDA assumes known, identifiable, mutually exclusive groups. Next, given group membership, the distribution of the exogenous variables is determined. That is, within each group assumptions are made about the distribution of the variables.

An advantage of using the logit model rather than MDA is that it is rel-

atively robust: many types of underlying assumptions lead to the same logistic formulation. No assumptions are made about the distribution of the independent variables, but only about the error terms. The linear MDA approach, by contrast, is applicable only if the underlying variables are multivariately normal with equal covariance matrices [Press & Wilson, 1978, p. 700]. When sampling from two multivariately normal populations with equal covariance matrices, both MDA and logistic regression can be used to derive valid estimates of the probability that a new observation should come from either one of the two populations. In that case, MDA produces asymptotically smaller relative classification error rates (*i.e.* the MDA estimator is asymptotically more efficient than the logit maximum likelihood estimator). When the assumptions for MDA are violated, the MDA estimator is not even consistent, whereas the logit (maximum likelihood) estimator is consistent and therefore more robust [Maddala, 1983, p. 27]. Even when conditions for MDA are met, the performance of logit is nearly as good as that of MDA for reasonable sample sizes. On the average, predicted probabilities from the two methods also show a small absolute difference when multivariate normality holds. Therefore, attention has slowly shifted from MDA to logit.

Applications

Martin [1977] was the first to use the logit approach to construct an early warning system to predict failure of financial institutions, banks in particular. The sample consisted of the entire population: 5575 nonfailed and 23 failed banks in 1975-1976, with financial variables from 1974. The model assigned a bank to the failed-bank group if its estimated probability was higher than the *a priori* probability of failure of 0.0041 (23/5598). Thus, the misclassification costs are assumed to be equal. The final model, which included four variables, correctly classified 91% of the nonfailed and 91% of the failed banks. The null hypothesis that all banks have equal probability of failure equaling 0.0041 was rejected for the logit model but not for the linear and the quadratic MDA models. However, when the different models were compared in terms of classification rather than probability estimation, the classifications of the logit and MDA models were virtually the same.

West [1985] applied factor analysis and logit estimation to early warning systems for commercial banks. Factor analysis was used to identify the common factors that influence the conditions of banks[4]. Each bank's score on a given factor is a normalized standard deviation and thus provides information on each bank with respect to the mean of the sample. The sample included 1900 banks and was

[4] The factors are composite variables that contain information distilled from a larger number of variables.

made up of banks with at least one examination for the period 1980-1982. Banks with CAMEL ratings[5] of 1 or 2 were considered sound and banks with ratings of 3,4, or 5 were classified as problem banks. In total, 19 variables were used, amalgamated into eight factors. The resulting model correctly classified about 90% of the banks that were used to build the system. No holdout sample was used.

In BarNiv [1990], logit analysis has been used to compare three different accounting principles for the property-liability (p-l) insurance industry. The purpose of this study was to determine which of the three accounting procedures, statutory accounting principles (SAP), generally accepted accounting principles (GAAP), or market value- and cash-flow-based principles (market value accounting MVA), provided better information for monitoring solvency and identifying financial distress. The sample consisted of 105 p-l insurers that failed in the period 1975-1987; these were matched with 106 nonfailed p-l insurers. Both MVA and SAP outperformed GAAP for all classification and validation comparisons. In a holdout sample, a five-variable model correctly classified 85%, 86%, and 72% of the insurers for SAP, MVA, and GAAP one year prior to bankruptcy. Three years before bankruptcy 76%, 83%, and 56% of the insurers were correctly classified for SAP, MVA, and GAAP.

Espahbodi [1991] developed a logit and an MDA model, which could identify potential failures in the banking industry. The original sample for this study consisted of 48 banks that failed in 1983, and another 48 non-failed banks that matched. For both models, four variables were selected based on a stepwise selection technique. The models were validated by predicting probabilities of failure in 1984. 1983 data were obtained on 76 (of a total of 79) banks that failed in 1984 and another 72 matching banks that did not fail. Using a cutoff point of 0.5, the overall classification accuracies of the logit and MDA models were 83% and 79%. A chi-square test showed that the difference was not significant (the p-value equaled 0.70).

Baranoff, Sager, and Witt [1993] applied cascaded logistic regressions to a disaggregated life and health (l&h) insurance industry. The l&h insurance industry is segregated into more homogeneous subgroups based on speciality and asset size. To each subgroup, a two-stage cascaded stepwise logistic regression methodology was applied. In the first stage, 256 solvency ratios for one company were

---

[5] CAMEL stands for Capital adequacy, Asset quality, Management, Earnings, and Liquidity. This system, which rates banks in the five areas above mentioned, was used by the three federal agencies that supervise and insure banks in the USA (*i.e.* the OCC, the Fed, and the FDIC). [West, 1985]

placed into one of 21 appropriate behavioral/financial categories such as underwriting, reinsurance, capitalization, etcetera. These 256 ratios were used as explanatory variables in the first-stage logistic regressions of each of the 21 categories. After this first stage, each company was described by a vector of 21 probabilities. Each of the 21 probabilities represented the ex ante probability that the company would experience difficulty, given the company's set of observed financial ratios in each category. The second stage involved a single logistic regression using the 21 probabilities to form a single summary measure of the risk of bankruptcy for the company concerned. The sample contained 49 bankrupt and 1750 nonbankrupt l&h insurers. Using the leaving-one-out technique, the models that emerged had a misclassification rate for bankrupt companies ranging from 0% for the most homogeneous subgroup to about 30% for the three least homogeneous subgroups.

Using logistic regression models, Cummins, Harrington, and Klein [1994] investigated possible improvements of the NAIC's property-liability Risk Based Capital (RBC) formula. The possible improvements the authors tested are: changes in the weights for the major components in the RBC formula, and incorporation of information on company size and organizational form. The authors compared RBC ratios from 1989 through 1991 for insurers who failed subsequently and insurers who survived through the first nine months of 1993. The samples consisted of 1567, 1606, and 1616 companies in 1989, 1990, and 1991 respectively. Seventy-four insurers with available data on 1989 failed subsequently, 51 with data on 1990, and 37 with data on 1991. The results indicated that less than half of the companies that failed eventually had RBC ratios within the proposed ranges for regulatory and company action. Thus, over 50% of the companies that later failed, did not fail to meet the RBC standard. Moreover, allowing the weights of the RBC components to vary and including information on firm size and organizational form improved the tradeoff between Type I error rates (the percentage of insurers that later failed and whose nonfailure was incorrectly predicted) and Type II error rates (the percentage of surviving insurers whose failure was wrongly predicted). Thus, the performance of the RBC standard can be improved by extending the standard with information of firm size and organizational form (stock or mutual) and by using other weights for the RBC components.

3.3.3 Other multivariate techniques

MDA and Logit are the most popular classification techniques. There are, however, some other techniques which have been applied successfully in this field.

Some studies have applied **Ordinary Least Squares** (OLS) for classification. Meyer and Pifer [1970] used OLS to predict bank failure. The sample consisted of 39 banks that closed between 1948 and 1965 and that were matched with 39 banks that remained solvent. The sample was randomly divided into an estimation sample of 30 pairs and a holdout sample of nine. The variables were selected by means of stepwise regression. The variables include financial ratios, trends, and variation and unexpected changes in financial ratios. The best predictions followed when six independent variables were included: all failed banks were correctly classified, and just one solvent bank was misclassified one year before failing. The number of correct classifications decreased slightly for data two year prior to failure. When the lead time was three years or more, financial variables were unable to discriminate between solvent and failing banks.

Eck [1982] used OLS to predict failure of property-liability insurers. Seven ratios were included in this model. By means of a stepwise selection technique, these ratios were selected from a group of seventeen ratios, which were selected on the basis of the assumption that most failures are the result of dishonesty. The sample included 25 firms which failed in the period 1965-1976. The 25 failed firms were matched to 25 nonfailed firms. The final model correctly classified 88% of a holdout sample of 60 companies (twelve failed), 92% of the nonfailed firms and 75% of the failed firms were classified correctly.

A problem with OLS is that the dependent variable is assumed to be continuous. For classification applications, the dependent variable is discrete (usually binary). Usually, the outcome of the model is interpreted as the ex ante probability of group membership (for example, bankruptcy) given X_n. However, the outcome may be beyond the zero-one interval. A probability that is negative or greater than one has no meaning. Furthermore, an outcome greater than one or less than zero must imply a violation of the assumption of zero-expectation errors, which is necessary for least squares estimation. The reader is referred to Altman *et al.* [1981] for a more extensive discussion. In Eck's model [1982], the outcome of the model ranged from -0.43 to 1.28. The author selected 0.5 as the cutoff point and did not discuss the interpretation of the outcomes.

Harrington and Nelson [1986] used OLS to estimate the relationship between premium-to-surplus ratios and property-liability insurer characteristics, including asset and product mix variables. In such case, the dependent variable is continuous and not restricted to the zero-one interval. According to the

authors, substantial deviations from the industry norm - in both directions - of the premium-to-surplus ratio given the asset and product mix may indicate higher or lower default probabilities than the average firm with similar characteristics. The estimation sample consisted of data for 1976 concerning 69 nonbankrupt firms. The model, which included 24 variables, was tested on a sample of data for 1976 concerning twelve firms, which went bankrupt between 1977 and 1981. The performance of the model on the testing set was compared with Best's rating and with the NAIC IRIS system; this led to mixed results.

Santomero and Vinso [1977] developed a basis to evaluate the cross-section riskiness of the banking industry, applying a **stochastic process modeling approach** to the population of continual weekly reporters to the Fed. Failure was defined as the bank's capital account becoming zero or negative, and the probability of this event was estimated by considering period-to-period changes in the capital account as a random variable whose distribution remains stationary over time. The parameters of this distribution for each bank can be estimated by analyzing weekly time series data reported by the bank and translated into a probability of failure. Although the distribution of changes in the capital account is influenced by profitability, loan losses, liquidity, and other factors, these elements were not explicitly included in the analysis. Rather, their influence is reflected in the estimated distribution of changes in the capital account. The authors used data concerning 224 banks for 1965 through early 1974, and concluded that the overall risk of failure in the banking system is extremely low.

BarNiv and Raveh [1989] presented a **Nonparametric Discriminant Model** (NPDM) to identify financial distress. The method differs from MDA in the separation rule applied; that is, a different quantity - called the index of separation - is maximized. The method searches for an optimal linear combination of the variables, which yields minimum overlapping between the scores given to group 1 and the scores of group 2. Contrary to MDA, no assumption of specific parametric (*e.g.* multivariate normal) distributions is needed. The model was applied to two samples. The first sample included 200 industrial firms, including 58 that failed between 1971 and 1981. The second sample included 69 non-life insurance companies that failed between 1975 and 1983, and 69 insurers that continued to be solvent. The empirical findings showed that NPDM outperformed MDA and logit analysis in terms of validation. In almost all cases better results were obtained for different costs and prior probabilities.

The logit model is a special case referring to a larger class of models, known as **generalized qualitative response models**. This class also includes the probit model, which has also been used for classification tasks, generalizations of the logit model (the lomit and burrit models), and - even more general - the expo-

nential generalized beta distribution of the second kind (EGB2). McDonald [1992, p. 228] argues that "the more general a model the better it should do relative to the log-likelihood value (a measure of fit, ELK). However, this does not imply an improvement in predictive ability, but the increased flexibility of the distribution certainly provides that potential". The author estimated MDA, probit, logit, and several generalizations of probit and logit to predict p-l insurer bankruptcy. The sample consisted of 212 observations (35 bankrupt and 177 nonbankrupt companies in 1983-1987). The qualitative response models appeared to be better able than MDA to predict both bankrupt and nonbankrupt companies. The qualitative response models yielded similar results, with the more general forms having marginally better predictive ability. Moreover, the models appeared to predict bankruptcy more accurately two years prior (between 96% for logit and 97% for EGB2) than one year prior to bankruptcy (94% correct for all models). This may be due to "window dressing" by the distressed companies.

BarNiv and McDonald [1992] also compared five different qualitative response models (Logit, Probit, Lomit, Burrit, and EGB2) to MDA and NPDM. They used a sample of 294 p-l insurers - 153 nonbankrupt and 141 bankrupt in 1974-1988 - one year prior to bankruptcy. A time series holdout sample was selected. This was done by estimating the models from data concerning companies from 1974 through 1983 for the estimation sample (83 nonbankrupt; 76 bankrupt) and by using different companies in a different time period (from 1984 through 1988) as an independent holdout sample (70 nonbankrupt and 65 bankruptt). Seven variables were included in the models. For the holdout sample, the seven models yielded similar results for all years prior to bankruptcy between 82% and 85% classified correctly one year prior to bankruptcy, between 79% and 80% two years prior to bankruptcy, and between 70% and 73% three years prior to bankruptcy. No model completely outperformed all other models.

Frydman, Altman, and Kao [1985] introduced **Recursive Partitioning** (RP) to predict financial distress for industrial firms. RP is a computerized, nonparametric classification technique, based on pattern recognition. It has attributes of both the classical univariate approach to classification and multivariate procedures. A sample is divided into two subsamples on the basis of the "best" splitting rule. Next, the process is repeated for each subsample; that is, each subsample is further divided according to the best splitting rule for the subsample. The best splitting rule is selected from the class of univariate splitting rules. Univariate splitting rules are rules that involve splitting an axis of one variable at one point. The models that result from RP are in the form of a binary classification tree, and assign objects into selected *a priori* groups. RP generally outperformed MDA on a sample of 200 firms, which included 58 bankrupt firms that failed between

1971-1981.

Carson [1994] applied MDA, logistic regression, and RP to detect financially distressed life insurers. The sample consisted of 40 insurers that did and 1380 insurers that did not go bankrupt in 1990 and 1991. The logit model misclassified the smallest number of nonbankrupt insurers (13%), whereas RP misclassified the least bankrupt insurers (17%). Compared to MDA, the logit model dominated in terms of the smallest number of misclassifications of nonbankrupt and bankrupt insurers. Overall, the logit model performed best with 87% of the insurers having been classified correctly.

Messier and Hansen [1988] used **Rule Induction** (RI), which can be seen as a special form of RP. RI attempts to discover regularities (rules) by analyzing a series of instances or examples related to a particular problem. Using RI, two expert systems were developed: one for the prediction of loan default and one for the prediction of bankruptcy of development firms. RI outperformed MDA with respect to both applications.

Kolari, Caputo, and Wagner [1994] examined the classification and predictive ability of **Trait Recognition** in comparison to MDA and logit analysis. Trait recognition (TR) is a pattern recognition technique that considers all possible interactions of the independent variables - taking them one, two, and three at a time -, and evaluates them for their potential usefulness in discrimination. TR has been applied to problems such as earthquake prediction, uranium detection, and drilling for oil. This is the first application in the field of business and economics; TR was applied to the prediction of commercial bank failure. The sample consisted of 126 commercial US banks that failed in 1986. The sample was completed with 878 randomly selected nonfailed banks. Twenty-eight variables were included in the models. TR considerably outperformed logit and MDA. The maximum error rate using TR in the holdout samples was 5%, which was less than the minimum error rate in the logit model in similar tests. The predictive ability of TR on holdout samples exceeded most results published on the general subject of predicting firm failure.

TR considers a large number of possible interactions among the independent variables, which may explain the strong performance of TR as compared to MDA and logit. Another pattern recognition technique that is also able to consider interactions among independent variables is the **Neural Network** technique. In contrast to TR, with only one known application, there have been quite a few successful applications of Neural Networks in the field of the classification of financial institutions. Neural network theory and a number of applications will be discussed in chapter 4.

3.4 POLYTOMOUS MULTIVARIATE RATIO ANALYSIS

It is not always possible to distinguish just two groups for a classification task. Even if it is possible to pool the outcomes into two groups only, it may not be desirable since relevant information may be lost. Therefore, it may be necessary to use a model specification that is able to handle a dependent variable that can take on more than two values. Many of the dichotomous models described in the previous sections can be extended to the polytomous case. However, the complexity of the models can grow dramatically by increasing the number of possible values of the dependent variable. In the following paragraphs, a short description will be given of the polytomous extensions of MDA, logit, and probit. The reader is again referred to Altman *et al.* [1981] for a more extensive description.

3.4.1 Multiple Discriminant Analysis

With reference to multiple discriminant analysis, the multiple group case is very similar to the two group case. We assume k discrete, mutually exclusive, identifiable populations or groups. Each group is assumed to contain observations that are characterized by a vector of m independent variables, X. Within each group, it is assumed that the observations of the independent variables are multivariate normally distributed. Thus, denoting X_{in} as the nth observation from the i-th group, with y_n as the variable indicating group membership,

$$X_{in} \sim N(\mu_i, \Sigma_i) \quad \text{given } y_n = i, \text{ for } i = 1, ..., k,$$

where $\mu_1, ..., \mu_k$ are m-length mean vectors and $\Sigma_1, ..., \Sigma_k$ are $(m \times m)$ covariance matrices.

The problem with classifying observations into one of k populations is often reduced to assigning an observation to the group with the highest probability. Under the assumption of equal group covariance matrices, *i.e.* $\Sigma = \Sigma_1 = \Sigma_2 = ... = \Sigma_k$, the probability of an observation belonging to group i equals

$$P(y_n = i \mid X_n) = \frac{\pi_i \exp(-0.5\,\chi_i^2)}{\sum_{j=1}^{k} \pi_j \exp(-0.5\,\chi_j^2)},$$

where

$$\chi_i^2 = (X_n - \mu_i)'\Sigma^{-1}(X_n - \mu_i) ,$$

and π_i the *a priori* probability of belonging to group i.

In case the researcher perceives costs of misclassification, and in case it is assumed that the costs of misclassifying an observation are the same regardless of which group it is assigned to, and group covariances are assumed to be equal, the classification rule becomes: assign an observation to group *i* if for all other groups *j*

$$X_n'\gamma_{ij} - \alpha_{ij} \geq \ln[C(-|j)\pi_j / C(-|i)\pi_i] ,$$

with C(-|*i*) being the cost of misclassifying an observation that actually belongs to group *i*, and with

$$\gamma_{ij} = \Sigma^{-1} (\mu_i - \mu_j) ,$$

and

$$\alpha_{ij} = (\mu_i + \mu_j)'\gamma_{ij} /2 .$$

Thus, this rule amounts to a series of paired two-group linear classification comparisons requiring only linear functions of X_n. This rule shows that the choice between any two groups *i* or *j* does not depend on the characteristics of any other group. In total, (*k* - 1) linear functions have to be calculated in order to apply this rule.

3.4.2 Logit Analysis

The logit and probit models can be classified into two broad, general categories. For the first type, it is assumed that there is a natural ranking in the possible values of the dependent variable. An example would be bond rating where the possible dependent variable values are **ordered** as Aaa is higher than Aa, which is higher than A, etcetera. For the second type, alternative dependent variables are assumed to be **unordered**. For example, a model to predict career choices of seniors at university. It would be difficult to rank these choices, *i.e.* say that doctors are ranked higher than lawyers. In the following paragraphs first the ordered

case will be discussed, followed by a discussion of the unordered case.

Suppose y_n can take on k **ordinally ranked values**, with $y_n = 1$ as the "lowest" and $y_n = k$ as the "highest" value. Furthermore, suppose

$$\begin{aligned} P_{1n} &= F(X_n'\beta) \\ P_{1n} + P_{2n} &= F(\alpha_2 + X_n'\beta) \\ &\vdots \\ P_{1n} + P_{2n} + \cdots + P_{k-1,n} &= F(\alpha_{k-1} + X_n'\beta) \\ P_{kn} &= 1 - F(\alpha_{k-1} + X_n'\beta) \end{aligned} ,$$

whith P_{in} as the conditional probability that choice i occurs for the n-th observation, X_n as a vector of m independent variables, β as an m-length parameter-vector,

$$\alpha_{k-1} > \alpha_{k-2} > \cdots > \alpha_2 > \alpha_1 = 0 ,$$

are threshold parameters, and

$$F(\alpha_i + X_n'\beta) = \frac{1}{1 + \exp[-\alpha_i - X_n'\beta]} ,$$

is the cumulative logistic function. This model is referred to as the **ordered logit model**.

The motivation behind this model (see McKelvey & Zavoina [1975]; Greene [1990, p. 703-704]) is that there is a continuous yet unobserved variable Y_n which is a linear function of the Xs and a stochastic standard logistic variable ϵ_n (μ=0, σ=1.81). Thus

$$Y_n = -X_n'\beta + \epsilon_n .$$

A standard normal ϵ_n would give rise to an **ordered probit model**, in which case the distribution of F( ) in the preceding paragraph is changed into the cumulative standard normal. In contrast to discriminant analysis, in which multivariate normality is assumed, with respect to the ordered logit and probit models no assumptions are made concerning the distribution of the independent variables.

The observed choice y_n is determined by the value of Y_n: $y_n = 1$ if $Y_n \leq 0$, $y_n = 2$ if $0 < Y_n \leq \alpha_2$, ..., $y_n = k - 1$ if $\alpha_{k-2} < Y_n \leq \alpha_{k-1}$, $y_n = k$ if $\alpha_{k-1} < Y_n$. The

width of each range is fixed across observations. However, the probability of Y_n falling in each range will vary with observations as $X_n'\beta$ changes. The scaling variables $\{\alpha_2, \ldots, \alpha_{k-1}\}$ are estimated by the data rather than imposed externally; that is, both the β-vector and the α-vector are estimated by maximum likelihood.

An example of an application of this type of model to finance is given by Kaplan and Urwitz [1979], who predicted Moody's bond ratings using an ordered probit model. They assumed the existence of an underlying continuous "bond risk" variable, which Moody's breaks into categories. Each bond is assigned the ordinal rating (Aaa, Aa, etcetera) of the category under which its risk comes. An inherent assumption of this model is that Moody's observes the ϵ_n's (hence Y_n) even though the researcher does not. A simple model using four variables could correctly classify about two-thirds of a holdout sample of newly issued bonds. With two additional financial ratios the prediction capacity can be improved a little further. Bonds were predicted never more than one rating category away from the one they actually came in.

Kim, Weistroffer, and Redmond [1993] used an ordered logit model to predict Standard & Poor's bond rating. For the training data, 110 companies from 1988 were randomly selected. Another 58 companies from the same year were randomly selected for the classification data. Finally, 60 companies from 1989 were randomly selected for the prediction data. Thus, two test sets were used, one from the same year as the training data and one from the following year. The ordered logit model correctly classified 43% of the classification set and 33% of the prediction set. Eighty-one per cent of the classification set and 75% of the prediction set were predicted into more than one rating category away from the one they actually came in. This model outperformed regression (OLS), MDA, and recursive partitioning (RP) for the classification set, and presented results similar to those of the neural network model. For the prediction set, OLS, MDA, RP, and ordered logit gave similar results, and the neural network model outperformed all of them.

With respect to many applications, there will be no natural ranking of alternatives (*i.e.* values are **unordered**). In such case, the **multinomial logit model** may be considered. This model has the following basic form: for each observation, assume that the dependent variable y_n can take on k discrete values and that the conditional probability of each choice i, given X_n (*i.e.* the probability $y_n = i$), is

$$P_{in} = P(i \mid X_n) = \frac{\exp(X_{in}' \beta_i)}{\sum_{j=1}^{k} \exp(X_{jn}' \beta_j)} .$$

This formulation differs from previous models in several respects. First, both the β vectors and the X variables are presented with subscripts denoting choices. This implies that both the X variables and their weightings may differ by choice. Putting it differently, the coefficient vector β for the comparison of choice i to choice j can be different from the coefficient vector for the comparison of choice i to choice k. Furthermore, different independent variables can be relevant in the comparison of choice i to choice j as in the comparison of choice i to choice k. According to Altman [1981, p. 71], almost all applications have taken the form of one of two different restricted versions. In the *multiple data model* (also referred to as the conditional logit model), it is assumed that there is just one coefficient vector, *i.e.* $\beta_i = \beta$ for all i. In the *multiple weighting model*, the βs are presented with subscripts denoting choices, but the Xs are not, *i.e.* $X_{in} = X_n$.

If the dependent variable can take on k different values, $(k - 1)$ logit functions are required. For instance, in the three-category model, category 1 can be taken as the reference outcome value and two logit functions are estimated: one for $y_n = 2$ versus $y_n = 1$, the other for $y_n = 3$ versus $y_n = 1$. The logit for comparing $y_n = 3$ to $y_n = 2$ may be obtained as the difference between the logit of $y_n = 3$ versus $y_n = 1$ and the logit of $y_n = 2$ versus $y_n = 1$. The two logit functions can be denoted as

$$g_2(X_n) = \ln[\frac{P_{2n}}{P_{1n}}] = X_n'\beta_2 ,$$

$$g_3(X_n) = \ln[\frac{P_{3n}}{P_{1n}}] = X_n'\beta_3 .$$

Schmidt and Strauss [1975] applied a five-category (multiple weighting) multinomial logit model to predict the occupation of individuals, using race, sex, educational accomplishment, and labor market experience as explanatory variables. Five occupational groups are distinguished: professional, white collar, craft, blue collar, and menial. Using menial as the reference group, four logit functions were estimated: the logit (*i.e.* the log of the odds) of blue collar against menial, craft against menial, etcetera. Race and sex were found to have had great effects; that is, among people of equal education and experience, race and sex strongly affect the type of job people will obtain.

3.5 SUMMARY

Logit and (linear) MDA are the most popular techniques to classify, among others, financial institutions. An advantage of using the logit model rather than MDA is that it is relatively robust: many types of underlying assumptions lead to the same logistic formulation. No assumptions are made about the distribution of the independent variables, but only about the error terms. The linear MDA approach, by contrast, is applicable only if the underlying variables are multivariately normal with equal covariance matrices. Even when the conditions for MDA are met, the performance of logit is nearly as good as that of MDA for reasonable sample sizes. Other relatively popular statistical classification methods are OLS and recursive partitioning.

In this study a polytomous extension to the standard binary logit model will be used, *i.e.* the ordered logit model. With the ordered logit model the dependent variable can take on a (finite) number of ordinally ranked values. The dependent variable used in this study, the risk exposure of a non-life insurance company, can take on three possible values: low, medium, and high. The model will be described in chapter 6.